PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-073415

(43)Date of publication of application : 16.03.1999

(51)Int.Cl. G06F 17/30

(21)Application number : 09-231363

(22)Date of filing : 27.08.1997

(71)Applicant : TOSHIBA CORP
TOSHIBA COMPUT ENG CORP

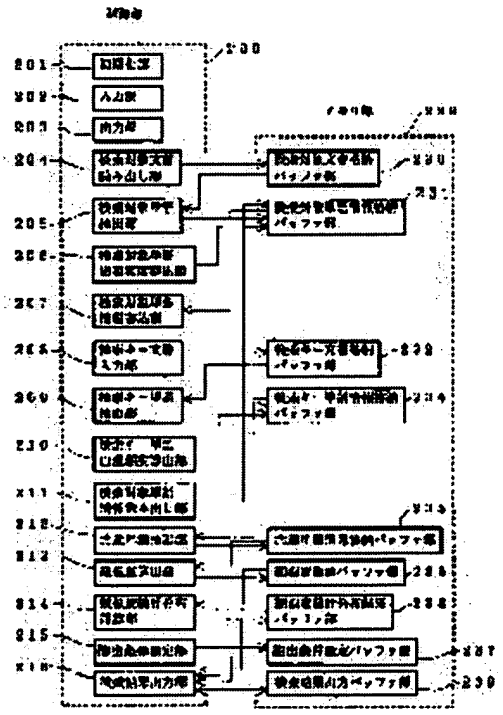
(72)Inventor : TANOSAKI YASUO
NAKAMOTO YUKIO
NISHINA TAKUYA
KUBOTA NAOHIDE

(54) DEVICE AND METHOD FOR RETRIEVING SIMILAR DOCUMENT

(57)Abstract:

PROBLEM TO BE SOLVED: To retrieve a similar document with high adequacy by finding statistical information on the similarity of each document to be retrieved and retrieving the similar document according to extraction conditions of the similar document set based on the similarity and statistical information.

SOLUTION: A common word extraction part 212 extracts the kind of a matching word and its appearance frequency information by comparing word information of a retrieval key document with word information of a document to be retrieved. A similarity calculation part 213 calculates the similarity between the retrieval key document and document to be retrieved according to the information. A similarity statistical distribution calculation part 214 finds statistical distribution calculation part 214 such as the mean value of similarity and standard deviation value. An extraction condition setting part 215 sets conditions for extracting the retrieval result of the similar document from the similarity statistical result. A retrieval result output part 216 judges whether or not there is the document to be retrieved and the corresponding document when there is the document to be retrieved from the



BEST AVAILABLE COPY

statistical distribution information, extraction condition values, and the similarity values of respective documents to be retrieved and outputs the retrieval result to a display device.

LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－73415

(43)公開日　平成11年(1999)３月16日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/403	３５０Ｃ

審査請求　未請求　請求項の数４　ＯＬ　（全 11 頁）

(21)出願番号　特願平9－231363

(22)出願日　平成９年(1997)８月27日

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(71)出願人　000221052
東芝コンピュータエンジニアリング株式会社
東京都青梅市新町３丁目３番地の１

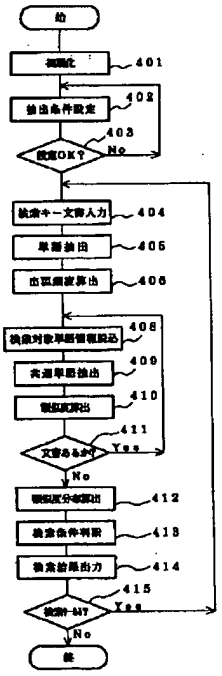
(72)発明者　田野崎　康雄
東京都青梅市末広町２丁目９番地　株式会社東芝青梅工場内

(74)代理人　弁理士　須山　佐一

最終頁に続く

(54)【発明の名称】　類似文書検索装置及び類似文書検索方法

(57)【要約】
【課題】　ある文書（検索キー文書）と類似する文書を複数の検索対象文書のなかから検索する装置において、より信憑性の高い類似文書検索を実現する。
【解決手段】　検索キー文書と各検索対象文書との各々の類似度値の統計分布（例えば類似度の平均値）を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出する。従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。また、検索キー文書と各検索対象文書との類似度がどれも一般的な評価基準において高いとは言えないような場合に、類似文書がないことを検索結果として出力する。



1

【特許請求の範囲】
【請求項１】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、
前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、
前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、
前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、
前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする類似文書検索装置。
【請求項２】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、
前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、
前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、
前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、
前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする類似文書検索装置。
【請求項３】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、
前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、
前記算出された各検索対象文書の類似度の統計情報を求める工程と、
前記統計情報を基準とする類似文書の抽出条件を設定する工程と、
前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする類似文書検索方法。
【請求項４】　検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、
前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、
前記算出された各検索対象文書の類似度の統計情報を求める工程と、
前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、
前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工程とを具備することを特徴とする類似文書検索方法。

2

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、電子化された文書データの検索装置に係り、特にある文書データを検索キーとしてこれと類似した文書データを自動検索する類似文書検索装置および類似文書検索方法に関する。
【０００２】
【従来の技術】近年、大量の電子化された文書データが流通するようになり、自動分類等を行う目的で、文書データベース中から指定された文書（以下、検索キー文書と呼ぶ。）に類似する文書の自動検索を行うシステムが実用されてきている。従来の類似文書検索システムでは、検索キー文書に含まれている単語と他の文書（以下、検索対象文書と呼ぶ。）に含まれている単語とを比較し、共通する単語の種類や出現回数・場所などからベクトル空間法により類似度を算出し、最も類似度の高い検索対象文書を検索結果として出力したり、類似度の高い文書から順に出力していた。
【０００３】
【発明が解決しようとする課題】従来の類似文書検索方式は、検索キー文書と文書データベース中の各検索対象文書との類似度を各々算出し、より類似度の高い文書を判定する、例えば最大類似度のものを検索結果として出力している。しかし、その検索結果は必ずしも妥当なものとは言えない。すなわち、上記従来の類似文書検索方式により得た類似度に基づく検索結果は、各検索対象文書に対して求めた各類似度の単純な大小比較により得た検索結果にすぎないため、例えば、検索キー文書と特に類似している検索対象文書が１つも存在しないような場合でも、その中で最も類似度の高い検索対象文書を類似文書として無条件に出力してしまう。
【０００４】本発明はこのような課題を解決するためのもので、複数の検索対象文書のなかから検索キー文書との類似が際立っているものを確実に検索することのできる類似文書検索装置および類似文書検索方法の提供を目的としている。
【０００５】また、本発明は、多くの一般的な類似評価の基準において類似していると呼べる類似文書のみを検索結果として得ることで、信頼性の向上を図ることのできる類似文書検索装置および類似文書検索方法の提供を目的としている。
【０００６】
【課題を解決するための手段】上記目的を達成するために、本発明の類似文書検索装置は、請求項１に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める

統計情報算出手段と、前記統計情報を基準とする類似文書の抽出条件を設定する抽出条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記抽出条件設定手段により設定された抽出条件に基づいて類似文書を検索する検索手段とを具備することを特徴とする。

【０００７】また、本発明の類似文書検索装置は、請求項２に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索装置において、前記検索キー文書と前記各検索対象文書との類似度を算出する類似度算出手段と、前記類似度算出手段によって算出された各検索対象文書の類似度の統計情報を求める統計情報算出手段と、前記統計情報を基準とする類似文書の有無の判定条件を設定する判定条件設定手段と、前記類似度算出手段によって算出された各検索対象文書の類似度および前記判定条件設定手段により設定された判定条件に基づいて類似文書の有無を判定する判定手段とを具備することを特徴とする。

【０００８】さらに、本発明の類似文書検索方法は、請求項３に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の抽出条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された抽出条件に基づいて類似文書を検索する工程とを具備することを特徴とする。

【０００９】さらに、本発明の類似文書検索方法は、請求項４に記載されるように、検索キー文書に類似する文書を複数の検索対象文書のなかから検索する類似文書検索方法において、前記検索キー文書と前記各検索対象文書との類似度を算出する工程と、前記算出された各検索対象文書の類似度の統計情報を求める工程と、前記統計情報を基準とする類似文書の有無の判定条件を設定する工程と、前記算出された各検索対象文書の類似度および前記設定された判定条件に基づいて類似文書の有無を判定する工程とを具備することを特徴とする。

【００１０】請求項１および請求項３の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として検索することができる。

【００１１】請求項２および請求項４の発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度がいずれも多くの一般的な評価基準において高いと言えないような場合に類似文書が存在しないとし、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【００１２】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して詳細に説明する。

【００１３】図１は本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図である。

【００１４】同図に示すように、この類似文書検索装置は、ＣＰＵおよびメモリなどから構成される制御装置１、キーボードなどの入力装置２、類似文書の検索結果などを表示する表示装置３、および文書データや類似文書検索のための各文書の単語情報などを格納する外部記憶装置４から構成される。

【００１５】図２に本類似文書検索装置における制御装置１の構成を示す。制御装置１は制御部２００とメモリ部２２９からなる。

【００１６】制御部２００は、初期化部２０１、入力部２０２、出力部２０３、検索対象文書読み出し部２０４、検索対象単語抽出部２０５、検索対象単語出現頻度算出部２０６、検索対象単語情報書込部２０７、検索キー文書入力部２０８、検索キー単語抽出部２０９、検索キー単語出現頻度算出部２１０、検索対象単語情報読み出し部２１１、共通単語抽出部２１２、類似度算出部２１３、類似度統計分布計算部２１４、抽出条件設定部２１５、検索結果出力部２１６などから構成される。　メモリ部２２９は、検索対象文書格納バッファ部２３０、検索対象単語情報格納バッファ部２３１、検索キー文書格納バッファ部２３２、検索キー単語情報格納バッファ部２３４、共通単語情報格納バッファ部２３５、類似度格納バッファ部２３６、抽出条件設定バッファ部２３７、類似度統計分布結果バッファ部２３８、検索結果出力バッファ部２３９などから構成される。

【００１７】初期化部２０１は、上記各バッファ部の初期化を行う。入力部２０２は、ユーザによる入力装置２からの検索キー文書や抽出条件の設定など各種設定の入力を行う。出力部２０３は、入力部２０２により入力された検索キー文書などの各種設定内容を表示装置３に出力する。

【００１８】検索対象文書読み出し部２０４は、外部記憶装置４に格納されている検索対象文書に関する情報を文書データベース化するために、文書データベース化すべき文書情報を外部記憶装置４から読み込み、検索対象文書格納バッファ部２３０に格納する。

【００１９】検索対象単語抽出部２０５は、検索対象文書格納バッファ部２３０に格納されている検索対象文書からの単語の切り出しを行う。そして、切り出した単語のなかからその文書内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索対象単語情報格納バッファ

部２３１に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【００２０】検索対象単語出現頻度算出部２０６は、検索対象単語抽出部２０５により抽出された個々のキー単語について、検索対象文書中での出現頻度を算出し、これを検索対象文書の単語情報として検索対象単語情報格納バッファ部２３１に格納する。

【００２１】検索対象単語情報書込部２０７は、検索対象単語情報格納バッファ部２３１に格納されている検索対象文書の単語情報を外部記憶装置４に格納する。

【００２２】検索キー文書入力部２０８は、入力装置２から入力された検索キー文書の情報を検索キー文書格納バッファ部２３２に格納する。

【００２３】検索キー単語抽出部２０９は、検索キー文書格納バッファ部２３２に格納されている検索キー文書からの単語切り出しを行う。そして、その文書の内容を表す上でキーとなる単語を抽出し、抽出した単語種を検索キー単語情報格納バッファ部２３４に格納する。単語の切り出しは形態素解析などにより行い、その文書の内容を表す上でキーとなる単語の単語種は品詞情報（例えば「名詞」や「サ変名詞」）を使って表現する。

【００２４】検索キー単語出現頻度算出部２１０は、検索キー単語抽出部２０９により抽出された個々のキー単語について、検索キー文書中での出現頻度を算出し、これを検索キー文書の単語情報として検索キー単語情報格納バッファ部２３４に格納する。

【００２５】検索対象単語情報読み出し部２１１は、外部記憶装置４に格納されいてる各検索対象文書の単語情報（単語の出現頻度情報）を１文書分ごとに呼び出し、検索対象単語情報格納バッファ部２３１に格納する。

【００２６】共通単語抽出部２１２は、検索キー単語情報格納バッファ部２３４に格納されている検索キー文書の単語情報と検索対象単語情報格納バッファ部２３１に格納されている検索対象文書の単語情報とを比較して、一致する単語の種類と出現頻度情報を共通単語情報格納バッファ部２３５に格納する。

【００２７】類似度算出部２１３は、共通単語情報格納バッファ部２３５に格納されている情報に基づき検索キー文書と検索対象文書との類似度を算出し、その類似度値を類似度格納バッファ部２３６に格納する。

【００２８】類似度統計分布計算部２１４は、類似度格納バッファ部２３６に格納されている検索キー文書と全検索対象文書との類似度値から類似度の平均値や標準偏差値などの統計分布情報を求めて類似度統計分布結果バッファ部２３８に格納する。抽出条件設定部２１５は、入力装置２を介してユーザより入力された、類似度統計結果から類似文書の検索結果を抽出する場合の条件、または類似度統計結果から検索キー文書との類似文書があるとするための条件、または検索キー文書との類似文書がないとするための条件などの抽出条件値を抽出条件設定バッファ部２３７に格納（設定）する。

【００２９】検索結果出力部２１６は、類似度統計分布結果バッファ部２３８に格納されている統計分布情報、抽出条件設定バッファ部２３７に格納されている抽出条件値、さらには類似度格納バッファ部２３６に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書検索結果として、検索対象文書の有無、検索対象文書が有る場合の該当文書を判断し、その検索結果を検索結果出力バッファ部２３９に格納し、そして検索結果出力バッファ部２３９の内容を表示装置３に出力する。

【００３０】次に、本実施形態の類似文書検索装置の動作を説明する。

【００３１】最初に、検索対象文書のデータベースの作成手順を図３、図５、図６により説明する。図３はその手順を示すフローチャートである。

【００３２】まず、初期化部２０１により全バッファ部の初期化を行う（ステップ３０１）。続いて検索対象文書読み出し部２０４が、外部記憶措置４から複数のテキスト文書を読み出し、検索対象文書格納バッファ部２３０に検索対象文書として格納する（ステップ３０２）。具体例として、例えば図５に示すような内容のテキスト文書を検索対象文書の一つとして格納したとする。

【００３３】次に、検索対象単語抽出部２０５が、検索対象文書格納バッファ部２３０に格納されている個々の検索対象文書について、形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索対象単語情報格納バッファ部２３１に格納する（ステップ３０３）。

【００３４】次に、検索対象単語出現頻度算出部２０６が、検索対象単語情報格納バッファ部２３１に格納されている検索対象文書のキー単語について、検索対象文書全体での出現頻度を算出し、その結果を検索対象単語情報格納バッファ部２３１に格納する（ステップ３０４）。図６に検索対象単語情報格納バッファ部２３１の格納例を示す。このバッファ部２３１において単語と頻度は対応付けて記述される。例えばキー単語「文書」が文書全体のなかで２回出現している場合は頻度として「２」が記述される。

【００３５】このようにして検索対象単語情報格納バッファ部２３１に格納された情報は、検索対象文書のデータベースとして外部記憶装置４に蓄積される（ステップ３０５）。

【００３６】この後、検索対象文書格納バッファ部２３０に文書データベース化前の検索対象文書が残っているかどうかを判断し（ステップ３０６）、他に検索対象文書があればステップ３０２に戻って、その新たな検索対

象文書についての前記同様の文書データベースの作成が行われる。他に検索対象文書がなければ本処理を終了する。

【００３７】次に、類似文書の検索手順を図４、図７乃至図１６により説明する。図４は類似文書検索手順を示すフローチャートである。

【００３８】まず、初期化部２０１により全バッファ部の初期化を行う（ステップ４０１）。続いて抽出条件設定部２１５が起動される。抽出条件設定部２１５は、入力装置２を通じてユーザより、

１．類似度統計結果から類似文書の検索結果を抽出する場合の条件

２．類似度統計結果から類似文書が存在しているとするための条件

３．類似度統計結果から類似文書が存在していないとするための条件

などの抽出条件値の入力を受け付けて抽出条件設定バッファ部２３７に格納（設定）する（ステップ４０２）。より具体的には、図７に抽出条件設定バッファ部２３７の格納例を示しているように、１．の条件として、「平均類似度の２倍以上の類似度を持つ検索対象文書を検索結果とする。」などが設定される。

【００３９】２．の条件として、「平均類似度の２倍以上の類似度を持つ検索対象文書がある場合」などが設定される。

【００４０】３．の条件として、「すべての検索対象文書の類似度が０．１以下である場合」などが設定される。

【００４１】これらの抽出条件値はユーザにより任意に決定される。

【００４２】この抽出条件の設定が完了したら（ステップ４０３）、検索キー文書入力部２０８が起動される。検索キー文書入力部２０８は、入力装置２を通じてユーザより検索キー文書の入力を受け付け、入力した検索キー文書の情報を検索キー文書格納バッファ部２３２に格納する（ステップ４０４）。具体例として、図８に示すような検索キー文書が入力されてバッファ部２３２に格納されたとする。

【００４３】次に、検索キー単語抽出部２０９が、検索キー文書格納バッファ部２３２に格納されている検索キー文書から形態素解析などによって単語の切り出しを行い、切り出した単語のなかから文書内容を表すキー単語を抽出し、そのキー単語の単語種（例えば品詞情報）を検索キー単語情報格納バッファ部２３４に格納する（ステップ４０５）。

【００４４】次に、検索キー単語出現頻度算出部２１０が、検索キー単語情報格納バッファ部２３４に格納されている個々のキー単語について、検索キー文書全体のなかでの出現頻度を算出し、これを検索キー単語情報格納バッファ部２３４にキー単語と対応付けて格納する（ステップ４０６）。図９に検索キー単語情報格納バッファ部２３４の格納例を示す。このバッファ部２３４においてキー単語と頻度は対応付けて記述され、例えばキー単語「今後」が２回出現している場合は頻度として「２」が記述される。

【００４５】次に、検索対象単語情報読み出し部２１１が、外部記憶装置４に格納されている各検索対象文書の単語情報を１文書ごとに読み込み、検索対象単語情報格納バッファ部２３１に書き込む（ステップ４０８）。

【００４６】この後、共通単語抽出部２１２が起動され、共通単語抽出部２１２は、検索対象単語情報格納バッファ部２３１と検索キー単語情報格納バッファ部２３４とに共通に格納されているキー単語を検出し、共通単語情報格納バッファ部２３５に格納する（ステップ４０９）。例えば、図１０に示すように、図６の検索対象単語情報格納バッファ部２３１と図９の検索キー単語情報格納バッファ部２３４に共通するキー単語として「画像」が検出され、このキー単語「画像」とその頻度情報「３」を共通単語情報格納バッファ部２３５に対応付けて格納する。

【００４７】次に、類似度算出部２１３が、共通単語情報格納バッファ部２３５に格納されている情報に基づき検索キー文書と検索対象文書との類似度をベクトル空間法などにより算出し、その類似度値を類似度格納バッファ部２３６に格納する（ステップ４１０）。例えば、図１１に示すような各検索対象文書ごとの類似度が類似度格納バッファ部２３６に格納される。

【００４８】すべての検索対象文書について類似度計算が完了すると（ステップ４１１）、類似度統計分布計算部２１４が起動する。類似度統計分布計算部２１４は、類似度格納バッファ部２３６に格納されている各検索対象文書の類似度の統計分布を算出し、その結果を類似度統計分布結果バッファ部２３８に格納する（ステップ４１２）。例えば、図１２に示すように、各検索対象文書の類似度の平均値「０．２５」を類似度の統計分布情報として求まる。

【００４９】次に、検索結果出力部２１６が、抽出条件設定バッファ部２３７に格納されている抽出条件値を判断し（ステップ４１３）、その抽出条件値、類似度統計分布結果バッファ部２３８に格納されている統計分布情報、さらには類似度格納バッファ部２３６に格納されている各検索対象文書の類似度値から、検索キー文書に対する類似文書の検索結果として、検索対象文書の有無や、検索対象文書が有る場合の該当文書を判断し、その結果を検索結果出力バッファ部２３９に格納する。　例えば、１．の条件（類似度統計結果から類似文書の検索結果を抽出する場合の条件）に対し、図１３に示すように、上記条件を満足する検索対象文書があれば、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５０】また、２．の条件（類似度統計結果から検

索キー文書との類似文書が存在しているとするための条件）に対し、該条件を満足している場合は、図１４に示すように、当該検索キー文書を類似文書有りの検索キー文書として、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５１】さらに、３．の条件（類似度統計結果から検索キー文書との類似文書が存在していないとするための条件）に対し、該条件を満足する場合は、図１５に示すように、当該検索キー文書を類似文書無しの検索キー文書として、そのＩＤを検索結果出力バッファ部２３９に格納する。

【００５２】検索結果出力部２１６は、検索結果出力バッファ部２３９の内容を、例えば図１６に示すような形式で表示装置３に出力する（ステップ４１４）。図１６の例では、例えば図１４に示すＩＤ「２」の類似文書有りの検索キー文書と類似する検索対象文書としてＩＤ「１」「４２」「５４」「３１４」などがあることを示している。また、２．と３．の各条件により、類似する検索対象文書がないと判断された検索キー文書に対しては類似文書がないことが表示される。

【００５３】この後、次の検索キー文書がある場合はステップ４０４に戻ってその検索キー文書を入力し、以降同様の処理を行う。次の検索キー文書がなければ本処理を終了する。

【００５４】なお、本動作例では、１つの検索キー文書に対する検索結果を得たところでこれを表示するようにしたが、すべての検索キー文書に対する検索結果を得た後、検索結果を見たい検索キー文書を指定すると、その検索結果が表示されるようにしてもよい。

【００５５】このように本実施形態の類似文書検索装置においては、検索キー文書と各検索対象文書との各々の類似度値の統計分布（この実施形態では類似度の平均値）を求め、この統計分布を基準に、ユーザが設定した条件を満足するものを類似文書として抽出することで、従来のように単に類似度値が高いものを類似文書として抽出する方式に比べ、類似文書としてより信憑性の高いものを検索結果として得ることができる。すなわち、本実施形態は、類似度の統計分布を基準としているので、検索キー文書との類似度がその他多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得られる。

【００５６】また、従来の方式では、検索キー文書と各検索対象文書との類似度が、どれも一般的な評価基準において高いとは言えないような場合でも検索結果として類似文書を無条件に出力してしまうが、本実施形態では、このような場合において類似文書がないことを検索結果として出力する。このような点からも、本実施形態の類似文書検索装置によれば、類似文書として信憑性の高い検索結果を得ることができ、類似文書検索効率を大幅に高めることができる。

【００５７】

【発明の効果】以上説明したように本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の抽出条件に基づいて類似文書を検索することで、類似文書としてより妥当性の高いもの、つまり類似度がその他の多くの検索対象文書に比べ際立って高い検索対象文書を類似文書として得ることができる。

【００５８】また、本発明によれば、各検索対象文書の類似度の統計情報を求め、各検索対象文書の類似度と、統計情報を基準に設定された類似文書の有無の判定条件に基づいて類似文書の有無を判定することで、検索キー文書と各検索対象文書との類似度がいずれも一般的な評価基準において高いと言えないような場合に類似文書が存在しないとし、一般的な評価基準において類似していると言える類似文書だけを検索結果として得ることができる。

【図面の簡単な説明】

【図１】本発明に係る一実施形態の類似文書検索装置のハードウェア構成を示す図

【図２】図１の類似文書検索装置における制御装置の機能ブロック図

【図３】検索対象文書のデータベースの作成手順を示す図

【図４】類似文書検索手順を示す図

【図５】検索対象文書の例

【図６】検索対象単語情報の格納例

【図７】抽出条件の設定例

【図８】検索キー文書の例

【図９】検索キー単語情報の例

【図１０】共通単語情報の例

【図１１】各検索対象文書の類似度の例

【図１２】類似度の平均値の例

【図１３】類似文書の検索結果の例

【図１４】類似文書の検索結果の例

【図１５】類似文書の検索結果の例

【図１６】類似文書の検索結果の出力例

【符号の説明】

２００……制御部
２１０……初期化部
２０２……入力部
２０３……出力部
２０４……検索対象文書読み出し部
２０５……検索対象単語抽出部
２０６……検索対象単語出現頻度算出部
２０７……検索対象単語情報書込部
２０８……検索キー文書入力部
２０９……検索キー単語抽出部
２１０……検索キー単語出現頻度算出部
２１１……検索対象単語情報読み出し部

11

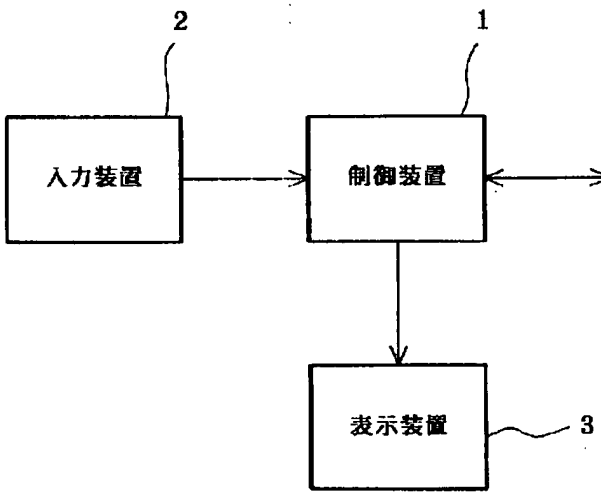
２１２……共通単語抽出部
２１３……類似度算出部
２１４……類似度統計分布計算部
２１５……抽出条件設定部
２１６……検索結果出力部
２２９……メモリ部
２３０……検索対象文書格納バッファ部
２３１……検索対象単語情報格納バッファ部

12

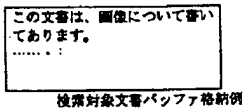
＊２３２……検索キー文書格納バッファ部
２３３……検索キー単語情報格納バッファ部
２３５……共通単語情報格納バッファ部
２３６……類似度格納バッファ部
２３７……抽出条件設定バッファ部
２３８……類似度統計分布結果バッファ部
２３９……検索結果出力バッファ部

＊

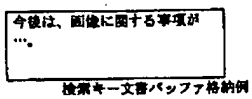
【図１】



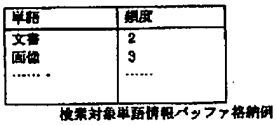
【図５】



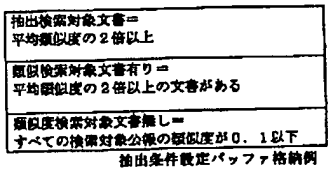
【図８】



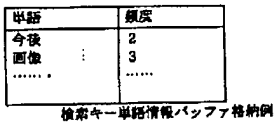
【図６】



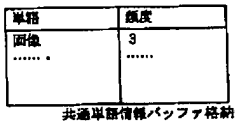
【図７】



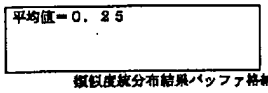
【図９】



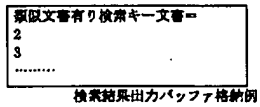
【図１０】



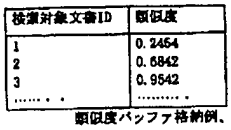
【図１２】



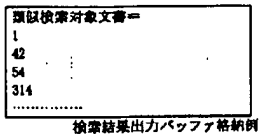
【図１４】



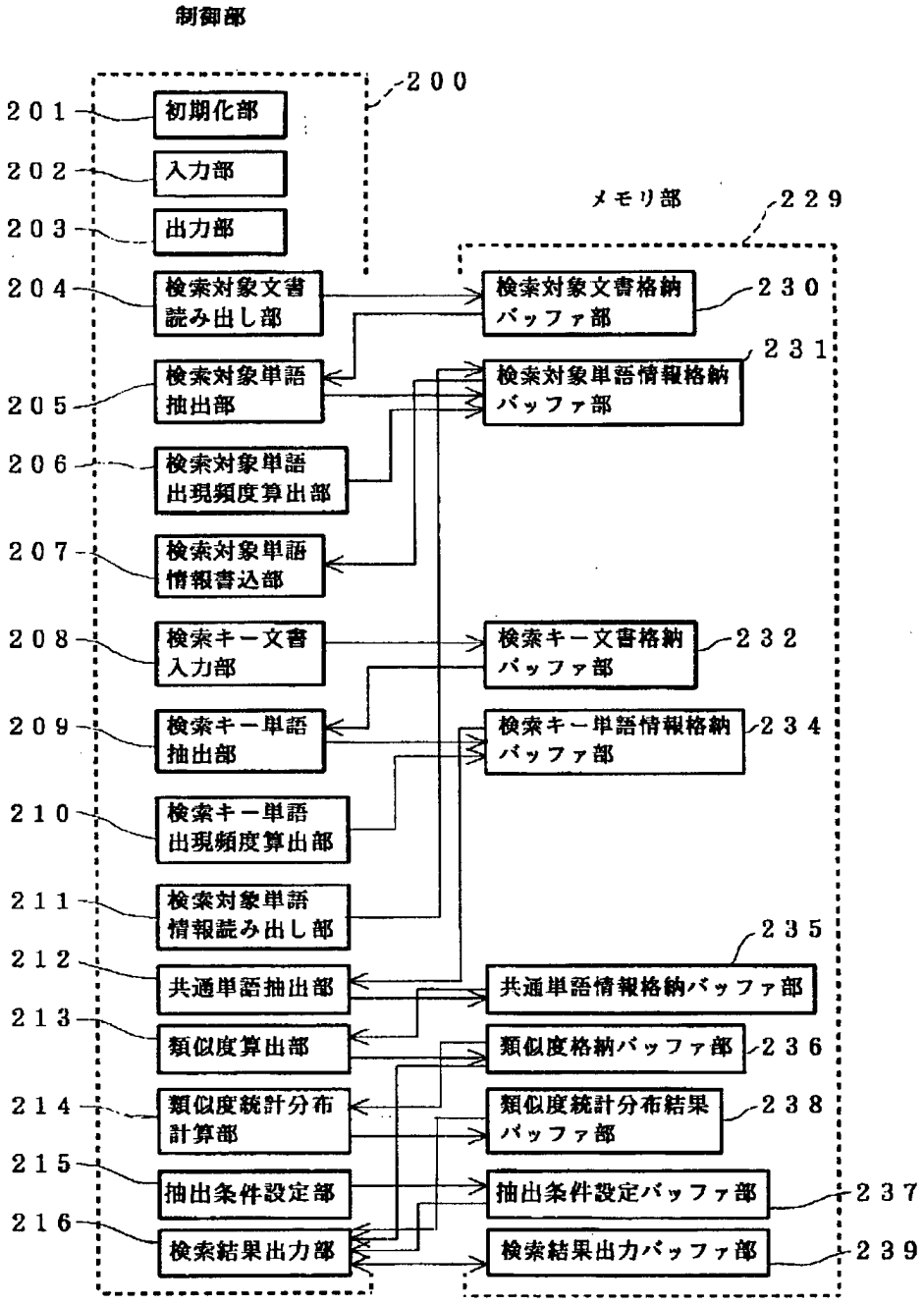
【図１１】



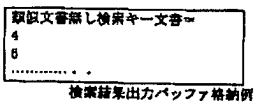
【図１３】



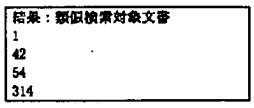
【図２】


制御部
２００
２０１
初期化部
２０２
入力部
２０３
出力部
２０４
検索対象文書読み出し部
２０５
検索対象単語抽出部
２０６
検索対象単語出現頻度算出部
２０７
検索対象単語情報書込部
２０８
検索キー文書入力部
２０９
検索キー単語抽出部
２１０
検索キー単語出現頻度算出部
２１１
検索対象単語情報読み出し部
２１２
共通単語抽出部
２１３
類似度算出部
２１４
類似度統計分布計算部
２１５
抽出条件設定部
２１６
検索結果出力部
メモリ部
２２９
検索対象文書格納バッファ部
２３０
検索対象単語情報格納バッファ部
２３１
検索キー文書格納バッファ部
２３２
検索キー単語情報格納バッファ部
２３４
共通単語情報格納バッファ部
２３５
類似度格納バッファ部
２３６
類似度統計分布結果バッファ部
２３８
抽出条件設定バッファ部
２３７
検索結果出力バッファ部
２３９

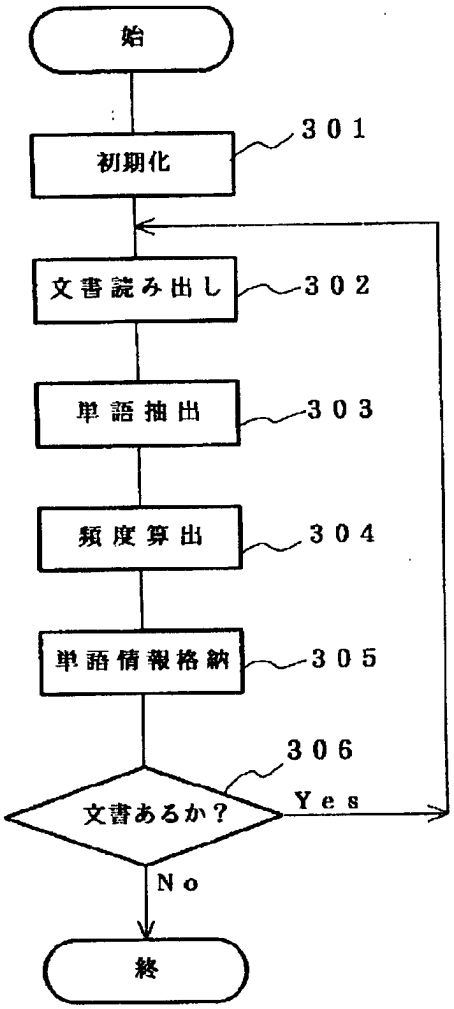
【図１５】


類似文書探し検索キー文書一
4
6
検索結果出力バッファ格納例

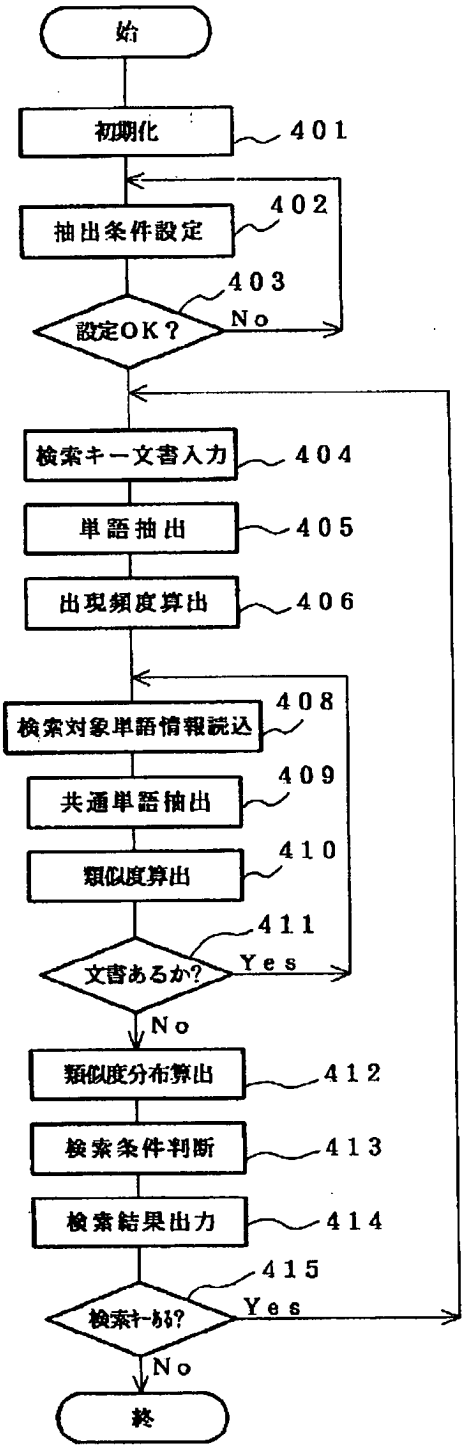
【図１６】


結果：類似検索対象文書
1
42
54
314

【図３】


始
初期化
３０１
文書読み出し
３０２
単語抽出
３０３
頻度算出
３０４
単語情報格納
３０５
３０６
文書あるか？
Ｙｅｓ
Ｎｏ
終

【図４】


始
初期化
401
抽出条件設定
402
403
設定OK？
No
検索キー文書入力
404
単語抽出
405
出現頻度算出
406
検索対象単語情報読込
408
共通単語抽出
409
類似度算出
410
411
文書あるか?
Yes
No
類似度分布算出
412
検索条件判断
413
検索結果出力
414
415
検索キーある?
Yes
No
終

フロントページの続き

(72)発明者　中本　幸夫
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　仁科　卓哉
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

(72)発明者　久保田　直秀
東京都青梅市新町1381番地１　東芝コンピュータエンジニアリング株式会社内

This Page is Inserted by IFW Indexing and Scanning Operations and is not part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☑ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.
As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.